4.7 Clock Generation for a 32nm Server Processor with Scalable Cores

Shenggao Li, Ashwin Krishnakumar, Edward Helder, Roan Nicholson, Vivian Jia

Intel, Santa Clara, CA

Within a given power envelope, the performance of a multi-core enterprise processor is greatly affected by inter-core (including I/O) data throughput and data transport latency. This paper presents the implementation of a clock system targeting low-power low-skew high-data throughput and low latency for a next-generation Xeon® server processor [1,2] with scalable cores in a 32nm 9-metal digital CMOS process.

Figure 4.7.1 is the die diagram of the processor featuring 8 SandyBridge cores [3], a shared last-level cache (LLC), a dedicated power control unit (PCU), and multiple high-speed I/O ports including Intel® QuickPath Interconnect (QPI) at 6.4GT/s or greater, PCI Express (PCIe) at 2.5/5GT/s, and DDR3 memory interface at 1600MT/s. The cores, cache, and I/O ports are attached through traffic-stops to an on-die ring-like bidirectional interconnect fabric running at the core clock rate. This highly modular ring architecture allows the number of cores and cache size to be scalable for various applications. The QPI link further offers the flexibility to form a multi-socket system [2]. The processor is powered by four major power domains, with VCCU serving the cores and caches, VCCP serving PCIe/QPI I/Os and PCU, VCCD serving the DDR3 I/Os, and VCCSA serving the system agents (i.e., PCIe/QPI/Memory logics). All power domain voltages (0.8 to 1.05V except VCCD, which is 1.35 to 1.5V) are controllable by PCU to allow voltage and clock frequency reduction or boost depending on thermal dynamics on the die. The clock architecture in Fig. 4.7.2 is designed to accommodate the design scalability, and power/clock domain crossing needs. The root of the clock system is an external 100MHz base clock (BCLK) providing reference clocks to the PCU, QPI, PCIe, and RCLK PLLs. The remaining PLLs are synchronized through the jitter filtering RCLK PLL, with a 100MHz reference clock (RxCLK) to the CPLLs/UPLL in the core and cache area, and a 133MHz reference clock (RyCLK) to the MPLLs in the memory system. Ideally, all PLLs and all clocks in the processor are synchronized to the same BCLK edge (defined as "Time0"). Clock jitter and path mismatches cause the end points in each clock tree to deviate from Time0. By taking into account the deviation, clock domains of same frequency can be crossed over synchronously using simple "de-skew" devices such as latch or D-flipflop. Clock domains of different frequencies are crossed over asynchronously using FIFO.

Data transport latency along the high-speed interconnect ring is a dominating factor on clock skew budgeting. In a 2-socket configuration with N cores per socket, every extra de-skew flop at each traffic-stop crossing causes a round-trip latency penalty of ~2(N+4) core clock cycles for an off-die data fetch. Assuming a non-scalable clock skew, the latency penalty gets worse when the clock cycle shrinks. Minimizing the inter-traffic-stop clock skew is the main motivation for adopting a single UPLL (see Fig. 4.7.1) for all the traffic stops. This is, however, at the expense of up to one cycle latency penalty at the core-cache crossings, due to the random UPLL-to-CPLL jitter (budgeted at 100ps). Skew reduction at the memory-agent and traffic-stop crossing is another design focus, by noting that RxCLK and RyCLK are distributed across a span of over 15×10mm^2. With matched clock trees between RxCLK and RyCLK, the estimated residual skew due to device mismatch and spatial voltage/temperature difference is still significant, and needs to be compensated using programmable delay lines. To ease the skew compensator design, the point of divergence (POD) on Rx|yCLK is pushed away from the RCLK PLL to the center of the die. Rx|yCLK is then retimed by a 400MHz clock at the POD. This reduces the skew range to be compensated by more than 50%.

Within each clock domain, one or more spine structure is used to deliver clocks across a large area to a sea of "clock islands", as shown in Fig. 4.7.3. Each clock island has a clock driver unit with a 1× and/or a 2× clock output (e.g., 800MHz and 1600MHz in the memory system) to serve a sub-region of about 200×200μm^2. The 1× clock is re-aligned to the 2× clock by synchronously resetting hundreds of distributed by-2 dividers. Static skew between the 1x and 2× clocks is negligible, in comparison with that of an alternative solution where the 1× and 2× clocks are distributed through two separate central spines. The islands are attached to the spine through balanced clock trees. The spine, the balanced trees, and the islands are individually tuned to minimize the global skew. Spines in the same clock domains are calibrated against each other using delay compensators with skew resolution of 5ps. An island distribution topology enables both skew and power reduction. It allows the fine tuning of pre-silicon clock skews to 10ps or below in each domain. Unused clock islands are disabled. A sparse clock grid and less metal utilization near each island helps in saving power.

To reduce clock jitter, a fifth power supply (1.8V) dedicated to PLLs and thermal sensors is routed through metal layers inside the package and down to the PLL power bumps directly. On-die voltage regulator is used to provide >20dB PSRR. All PLLs except the PCIe/QPI PLLs are based on a 3-stage ring-oscillator optimized for jitter. The oscillator is tuned by a capacitor array coarsely and a MOS varactor finely at each delay stage. To meet the tight jitter budget of the PCIe/QPI IO, LC-tank based VCO is chosen for the PCIe/QPI PLLs. The inductor is a 3-turn symmetrical structure using top level metals (M7-8, Q = 5, L = 0.8nH). It is the first time on-chip planar inductors are being used in an Intel® server processor. Being an indispensible device for RF and high-speed ICs, the inductors are yet incompatible with low cost digital CMOS process. Trade-offs have to be made with regard to fitting an inductor in a sea of package bumps with degraded performance, versus depopulating bumps around the inductor that may cause non-uniformity on packaging stress. Power grid discontinuity, routing congestion, and metal density non-uniformity are of other concerns since metal is blocked below and near inductors. Figure 4.7.4 illustrates the PCIe/QPI IO clock system. The LC-VCO is AC-coupled to two pseudo-differential CMOS buffers to drive a CMOS by-2 (or by-4) divider with 4-phase (or IQ phase) outputs needed for data recovery at ½ or ¼ of the VCO rate. The IQ clocks are distributed differentially using 2 stages of AC-coupled current-mode-logic (CML) buffers to drive 5mm on-die passive transmission lines [4]. All transceiver lanes are tapped to the T-lines. It is noted that precision resistors are not available in this process. NWELL based resistive loads on CML logic circuits have to be calibrated against an external 50Ω resistor. The bandwidth of the NWELL resistors is limited by parasitics. It is a challenge to design the post VCO buffer and IQ generators in CML logic to operate near and beyond 10GHz. The circular ring structures in Fig. 4.7.4 and Fig. 4.7.5 offer a solution to 4-phase (by-2 divider) and 8-phase (by-4 divider) generation. The back-to-back connected inverters are used for duty-cycle correction, but also prevent the ring from entering a "dead-lock" state. Figure 4.7.5 presents the state transition table of the ring emerging from any initial state and resolving to a desired phase relationship. The simplicity of this circuit allows it to operate at speed higher than any other non-regenerative divider structure, in addition to providing IQ phase and 50% duty cycle.

Figure 4.7.6 shows the measured phase noise of the LC PLL, at the aforementioned by-2 divider and by-4 divider outputs, with a jitter of 1ps$_{rms}$ integrated from 100Hz to 1GHz. Long term rms jitter for ring-oscillator based PLLs is less than 2ps, measured through a shared debug port. Time-resolved emission (TRE) is used for clock skew probing. A maximum skew of 14ps is obtained for a group of representative clock nodes in the cache area. The maximum compensator offset from its nominal setting is 2-LSB (1 LSB = 5ps) collected from delay compensators in the cache area.

Acknowledgment:
The authors thank Ernest Knoll, Yair Talker, Stephane Smith, Rohit Mittal, Luke Tong, and ST Chen for their contributions, Simon Tam, Stefan Rusu, and Fulvio Spagna for valuable inputs, and Abhi Kolla for managerial support.

References:
[1] Stefan Rusu, Simon Tam, et al, "A 45nm 8-core Enterprise Xeon Processor", *ISSCC Dig. Tech. Papers*, pp. 56-57, Feb. 2009.
[2] Nasser Kurd, et al, "Westmere: A family of 32nm IA Processors", *ISSCC Dig. Tech. Papers*, pp. 96-97, Feb 2010.
[3] Marcelo Yuffe, Ernest Knoll, et al, "A Fully Integrated Multi-CPU, GPU and Memory Controller 32nm Processor", *ISSCC Dig. Tech. Papers*, Feb. 2011.
[4] G. Balamurugan, et al, "A Scalable 5–15 Gbps, 14–75 mW Low-Power I/O Transceiver in 65 nm CMOS", *IEEE J. Solid State Circuits*, Vol. 43, No. 4, pp 1010-1018, Apr. 2008.

 978-1-61284-302-5/11/$26.00

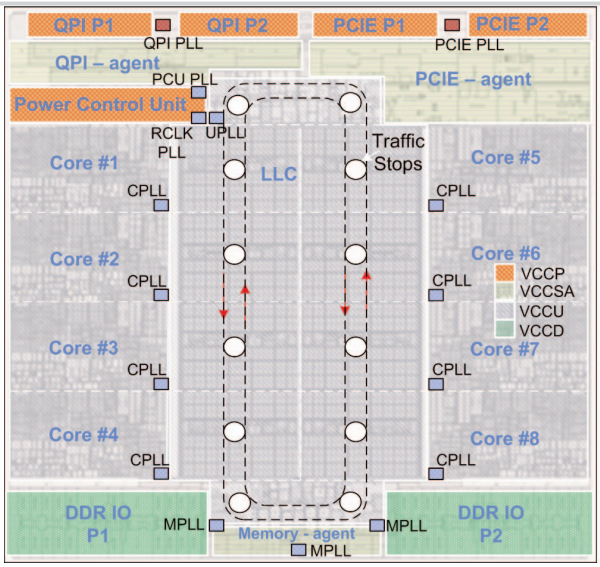

Figure 4.7.1: Die micrograph of the processor.

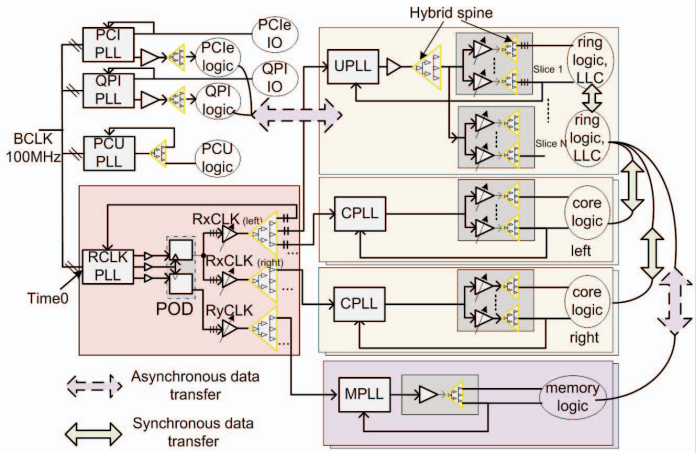

Figure 4.7.2: PLL and clock distribution architecture.

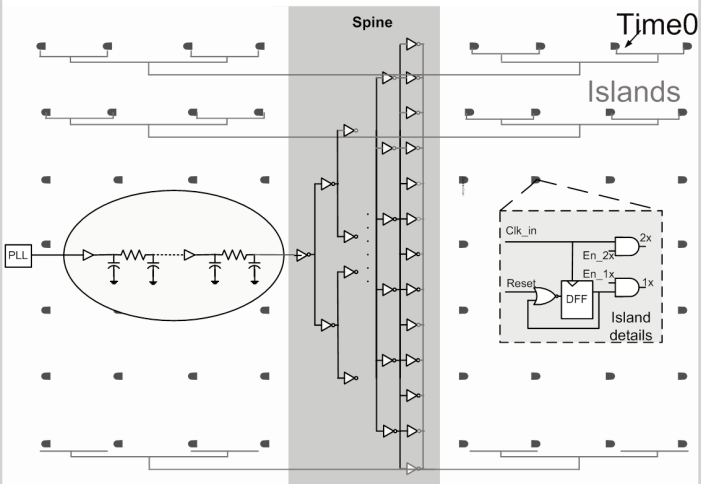

Figure 4.7.3: Islands based clock distribution.

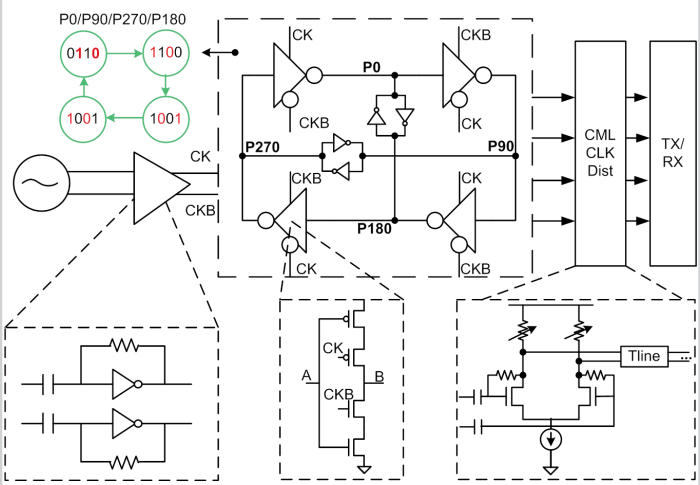

Figure 4.7.4: PCIe/QPI IO clock generation and distribution.

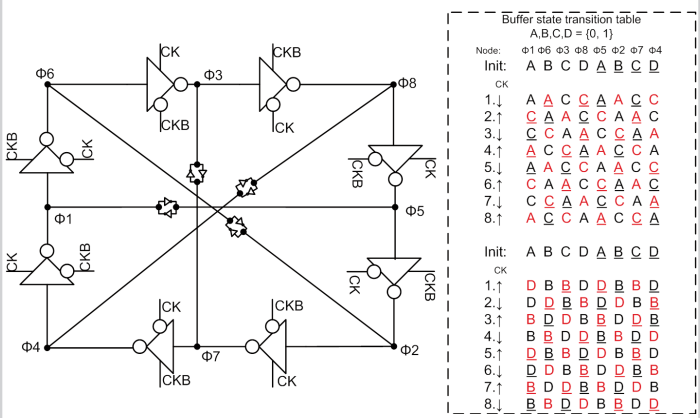

Figure 4.7.5: A by-4 divider with 8 phase generation.

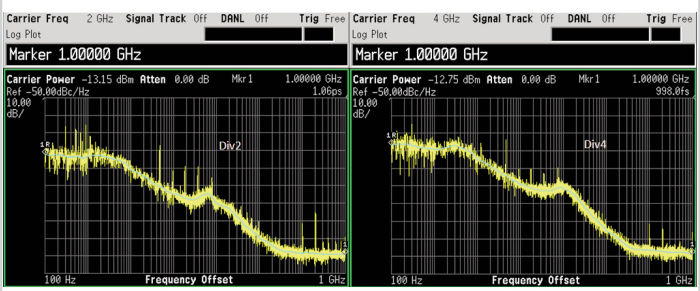

Figure 4.7.6: LC PLL Phase noise measurement at by-2, by-4 divider outputs.

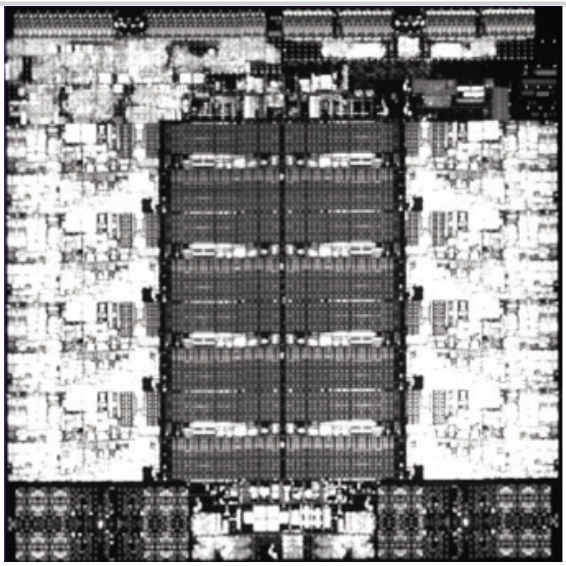

Figure 4.7.7: Die Photo.

978-1-61284-302-5/11/$26.00